12th International Conference on Information Fusion
Seattle, WA, USA, July 6-9, 2009

Information Theoretic Measures for Performance Evaluation and Comparison

Huimin Chen
Department of Electrical Engineering
University of New Orleans
New Orleans, LA 70148, U.S.A.
hchen2@uno.edu

Genshe Chen
DCM Research Resources, LLC
14163 Furlong Way
Germantown, MD 20874, U.S.A.
gchen@dcmresearchresources.com

Erik P. Blasch
AFRL/RYAA
Wright-Patterson AFB, OH 45433, U.S.A.
Erik.Blasch@wpafb.af.mil

Philip Douville
AFRL/RYAA
Wright-Patterson AFB, OH 45433, U.S.A.
Philip.Douville@wpafb.af.mil

Khanh Pham
AFRL/RVSV
Kirtland AFB, NM 87117, U.S.A.
Khanh.Pham@kirtland.af.mil

Abstract – *This paper discusses the performance comparison of different algorithms for classification, estimation and filtering problems. Two information theoretic measures, namely, the empirical mutual information and the asymptotic information rate are proposed for simulation based performance evaluation and algorithm comparison. They can be used as a guideline for designing a practical procedure to measure the performance of different algorithms with limited computational resources. Other useful performance measures are reviewed and their relation to the two new measures discussed. Several practical examples are used to provide some insights on the inherent difficulty of algorithm ranking and the advantage of using the information theoretic measures for algorithm comparison.*

Keywords: Performance evaluation, information theoretic measure, detection, estimation, filtering.

1 Introduction

Performance evaluation aims to study the behavior of a system operated by various algorithms and compare their pros and cons based on a set of measures or metrics each of which usually maps different algorithms into different real values or partial orders for ranking. In practical applications, as the system being studied becomes more and more complex and complicated, the analytical results regarding the performance of different algorithms with respect to a particular measure usually do not have closed forms or they are computationally intractable. Thus simulation based performance evaluation serves as an indispensable tool to measure the performance of various algorithms. On the other hand, there are several distinctive issues on how to develop a good procedure to collect and disseminate information from the system relevant to performance aspects. Good measures to reliably rank different algorithms as well as the assessment on the credibility of the ranking can also guide the algorithm development especially with limited computational resources. The issues related to performance evaluation metrics and algorithm development to optimize them are often interrelated and demand an interdisciplinary research on system design, data analysis, simulation methods, and statistical inference. In this paper, we do not intend to address the design issues for performance evaluation. Our major focus is on the performance evaluation of classification, estimation and filtering problems. The algorithms we want to compare are the so-called classifiers, estimators, filters or a combination of the above for joint problems. A natural measure is the classification error, estimation error or filtering error which requires some clarification for different types of problems. With this setting, one also seeks to develop an algorithm that achieves the minimum error under a given performance measure. Different error measures will result in different solutions each of which corresponds to the minimizer of a particular error measure. However, there is no consensus on which measure is particularly good for algorithm comparison. One often has to evaluate several measures and display all of the ranking results for a practitioner to make a decision based on a certain weighted combination of different measures in a common metric space.

We want to make a distinction between the algo-

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUL 2009		**06-07-2009 to 09-07-2009**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Information Theoretic Measures for Performance Evaluation and Comparison	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Department of Electrical Engineering,University of New Orleans, ,New Orleans,LA,70148**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM002299. Presented at the International Conference on Information Fusion (12th) (Fusion 2009). Held in Seattle, Washington, on 6-9 July 2009. U.S. Government or Federal Rights License.

14. ABSTRACT
This paper discusses the performance comparison of different algorithms for classification, estimation and filtering problems. Two information theoretic measures, namely, the empirical mutual information and the asymptotic information rate are proposed for simulation based performance evaluation and algorithm comparison. They can be used as a guideline for designing a practical procedure to measure the performance of different algorithms with limited computational resources. Other useful performance measures are reviewed and their relation to the two new measures discussed. Several practical examples are used to provide some insights on the inherent difficulty of algorithm ranking and the advantage of using the information theoretic measures for algorithm comparison.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Public Release**	**8**	

rithm developer (AD) and the performance evaluator (PE) since they may use different measures for different purposes. Furthermore, we assume that the problem being studied has a probabilistic structure and each algorithm provides the statistical inference of this probabilistic model so that the PE will gain certain amount of information by running an algorithm. With this problem formulation, we propose two information theoretic measures for ranking different algorithms. One measure is called the *empirical mutual information* between the PE and the AD, which depends on the size of the test data. To have an information theoretic measure which is independent of the data length, we propose another measure called *asymptotic information rate*, which characterizes the performance improvement of an algorithm with large data size for performance evaluation. These two measures are useful and complementary to some standard and existing error measures, some of which may not be suitable for the performance evaluation of joint classification, estimation, and filtering problems. We relate the two information theoretic measures to classical hypothesis testing, classification, quantization and estimation problems and provide insights on their usage in some nonstandard settings. Finally, we give several examples to show the usefulness of the information theoretic measures for algorithm comparison.

2 Performance Comparison Using Mutual Information Based Measure

In this section, we formulate performance evaluation as a statistical inference problem where the PE holds certain prior knowledge of the system parameters and the algorithm being evaluated provides additional information to reduce the uncertainty of those parameters. The mutual information quantifies how much the PE can gain by running a particular algorithm.

The problems being considered include hypothesis testing, classification, parameter estimation and filtering. We assume that the PE can generate data repeatedly based on the same or different probabilistic structures to test different algorithms with a predefined measure. The underlying truth that governs the data generation can be (1) a statistical hypothesis among many predetermined ones or (2) a value in the parameter space or (3) a time function from the realization of a random process. The algorithm is supposed to choose a hypothesis, produce an estimate of the unknown parameter or a time function of the unknown process based on the data generated by the PE. The purpose of the PE is to rank different algorithms using a set of measures. A fundamental question is what measures the PE tends to adopt. The measures have to be applicable to all the aforementioned problems and possibly joint ones.

To begin with, we consider a classification problem where the PE has K hypotheses denoted as $\{H_1, ..., H_K\}$. Each hypothesis can generate a data sequence $\{z_1, z_2, ...\}$ with a certain mechanism conditioned on that particular hypothesis being true. The data sequence is used by the algorithm to determine which hypothesis governs the underlying data generation mechanism. A standard measure is the classification error for each hypothesis. However, one can not directly compare the performance of two algorithms by looking at two arrays of the classification errors for all hypotheses. Of course one can use the weighted classification error as a measure and those weights can be derived by minimizing a Bayesian risk function if one has the prior probability of each hypothesis being used to generate the data [9]. In our formulation, these prior probabilities are determined by the PE and they are inherently subjective. However, the choice of probabilities also represents the PE's belief on which hypothesis is true without evaluating any algorithm.

Denote by X the discrete random variable with probability mass function (pmf) $P(X)$ given by

$$\Pr(X = H_i) = p_i, \quad i = 1, ..., K \tag{1}$$

For a data sequence $\{z_1, ..., z_N\}$ of length N, denote $\hat{X}_N$ the discrete random variable with pmf indicating the probability that a particular hypothesis is chosen by the classifier based on the data sequence of length N. The probability that hypothesis i is true conditioned on hypothesis j being chosen by the classifier is denoted by

$$\Pr(X = H_i | \hat{X}_N = H_j) = q_{ij}, \quad i = 1, ..., K; j = 1, ..., K \tag{2}$$

The entropy of X is

$$H(X) = -\sum_{i=1}^{K} p_i \log p_i \tag{3}$$

The conditional entropy of X given $\hat{X}_N$ is

$$H(X|\hat{X}_N) = -\sum_{i=1}^{K} p_i \sum_{j=1}^{K} q_{ij} \log q_{ij} \tag{4}$$

The mutual information between X and $\hat{X}_N$ is defined as

$$I(X; \hat{X}_N) = H(X) - H(X|\hat{X}_N) \tag{5}$$

It quantifies the reduction of uncertainty about which hypothesis is true based on the classification results. If the classifier always chooses the correct hypothesis, then the mutual information achieves its maximum $H(X)$. Note that if the classifier always chooses the incorrect hypothesis when testing two hypotheses, we still get the maximal mutual information and one can easily modify the decision of this classifier to achieve zero error. Thus the mutual information is a good indication

of the classification performance. If one wants to maximize the mutual information over the distribution of X, then a uniform distribution among the K hypotheses should be used to generate the data sequences, i.e., all classes have equal prior probabilities [6].

Another popularly used performance measure to rank classifier is the classification error. However, this measure can be misleading as illustrated in [14] (p. 532). To take the advantage of using both mutual information and classification accuracy in ranking the classifiers, one may consider a new measure, namely, the *normalized error to information ratio*, given by

$$\alpha_N \frac{e_N}{I(X;\hat{X}_N)} \tag{6}$$

where e_N is the classification error and α_N is a prespecified parameter based on the PE's preference between having a better classification accuracy and gaining more information. The best classifier should minimize the above measure. Note that when no error is made by the classifier, the mutual information gained by the performance evaluator becomes irrelevant. This is reasonable since the PE will gain the maximum information when the classifier has zero error no matter what prior distribution among the K hypotheses that the PE uses.

Next, we consider a parameter estimation problem where θ is unknown and to be estimated in a parameter space Θ. We use a vector $\mathbf{z}_N$ to denote the observed data sequence $\{z_1, ..., z_N\}$ of length N. The estimator uses $\mathbf{z}_N$ to estimate θ and provides the estimate $\hat{\theta}_N$. We assume that the estimate is in another space $\hat{\Theta}$. The performance evaluator has prior uncertainty about θ which is characterized by the probability density function (pdf) $f(\theta)$. The differential entropy of θ is

$$h(\theta) = -\int_{\Theta} f(\theta) \log f(\theta) d\theta \tag{7}$$

The differential entropy of θ given $\hat{\theta}_N$ is

$$h(\theta|\hat{\theta}_N) = -\int_{\Theta} f(\theta|\hat{\theta}_N) \log f(\theta|\hat{\theta}_N) d\theta \tag{8}$$

In practice, the PE can only concentrate on a parameter space of finite support especially when the likelihood function $\Lambda(\mathbf{z}_N|\theta)$ does not have a parametric form. For convenience, we assume that the prior pdf is proper and the above differential entropies always exist. In this case, the mutual information between θ and $\hat{\theta}_N$ is defined as

$$I(\theta;\hat{\theta}_N) = h(\theta) - h(\theta|\hat{\theta}_N) \tag{9}$$

Similarly, for a random process $\theta(t)$, an algorithm should provide the estimate $\hat{\theta}_N(t)$ and we can define the average entropy of $\theta(t)$ as

$$h^*(\theta(t)) = \lim_{t\to\infty} \frac{\int_{-\infty}^{t} h(\theta(u))du}{t} \tag{10}$$

The average mutual information between $\theta(t)$ and $\hat{\theta}_N(t)$ is defined as

$$I(\theta(t);\hat{\theta}_N(t)) = h^*(\theta(t)) - h^*(\theta(t)|\hat{\theta}_N(t)) \tag{11}$$

The major issue of the above generalization of the mutual information measure from hypothesis testing and classification problems to estimation and filtering problems is that the PE needs to evaluate the integral accurately when applying the mutual information measure which requires the knowledge of the continuous pdf $f(\theta|\hat{\theta}_N)$ or $f(\theta(t)|\hat{\theta}_N(t))$. In practice, estimating a continuous pdf using a finite number of realizations is an ill-posed problem and one has to assume certain properties (e.g., smoothness, finite dimensional parametric family) of the pdf in order to obtain a unique solution. Alternatively, if we modify the estimation and filtering problems so that the original probability measure is approximated by another measure defined in a finite partition of the parameter space, then the PE only needs to evaluate the pmf of θ or $\theta(t)$ in a finite discrete space, which makes the algorithm comparison feasible via computer simulations. We will explain this idea next.

3 Empirical Mutual Information and Asymptotic Information Rate

In this section, we convert the performance evaluation of a class of estimation and filtering problems into a classification problem with tolerable distortion of the mutual information based on a properly chosen distortion metric. We call the resulting measure *empirical mutual information* which can be applied to algorithm comparison without the ill-posed issue in density estimation.

A distortion measure is a mapping $d : \Theta \times \hat{\Theta} \to R^+$ from the set of parameter-estimate pairs into the set of nonnegative real numbers. A commonly used distortion measure is the squared error given by

$$d(\theta, \hat{\theta}) = ||\theta - \hat{\theta}||^2 \tag{12}$$

Note that the Euclidian distance between the parameter and its estimate is also a metric since the parameter space is usually a metric space. A distortion measure is said to be bounded if

$$\max_{\theta\in\Theta, \hat{\theta}\in\hat{\Theta}} d(\theta, \hat{\theta}) < \infty \tag{13}$$

In most cases, the parameter space and the estimate space are the same. We are interested in the partition of the parameter space with a desired distortion bound. An M partition of Θ, i.e., $\Theta = \Theta_1 \cup \cdots \cup \Theta_M$ and $\Theta_i \cap \Theta_j = \phi$, $\forall i \neq j$, is said to be d_M bounded if

$$\max_{\theta_0, \theta_1 \in \Theta_i} d(\theta_0, \theta_1) \leq d_M, \quad i = 1, ..., M \tag{14}$$

For a bounded parameter space and a bounded distortion measure, one can choose an appropriate value d_M to have a finite partition of the parameter space being d_M bounded. A particular subset Θ_i of the parameter space represents a hypothesis i while $\hat{\theta} \in \Theta_i$ represents the estimator chooses the correct hypothesis when $\theta \in \Theta_i$. A good distortion measure often has the property that d_M can be made arbitrarily small even when M is finite. The partition of the parameter space and the associated distortion measure are commonly treated in quantization theory for designing the best partition of the parameter space and the presentative value of each region to minimize the expected conditional distortion measure [7, 6]. Here our focus is to convert the estimation and filtering problem into a classification problem so that the mutual information measure can be evaluated to compare the performance among different estimators. One may argue that there exist several good performance measures for the estimation problem such as the mean square error, efficiency, consistency and unbiasedness [15, 1]. However, these measures require that the PE has the knowledge of the likelihood function of the data sequence being generated or the prior distribution of the parameter which may not be available for some practical problems.

We assume that, for any given value $\theta \in \Theta$, the PE can generate a data sequence of length N and evaluate the estimator using $\hat{\theta}_N$. For any subset Θ_i of the M partition of Θ, the distortion is d_M bounded. The prior distribution of θ being in one of the M partitions is given by

$$\Pr(\theta \in \Theta_i) = p_i, \;\; i = 1, ..., M \tag{15}$$

The conditional probability $\Pr(\theta \in \Theta_i | \hat{\theta}_N \in \Theta_j)$ is given by

$$\Pr(\theta \in \Theta_i | \hat{\theta}_N \in \Theta_j) = q_{ij}, \;\; i = 1, ..., M; j = 1, ..., M \tag{16}$$

The empirical entropy $H(\theta)$ is given by

$$H(\theta) = -\sum_{i=1}^{M} p_i \log p_i \tag{17}$$

The empirical conditional entropy of θ given $\hat{\theta}_N$ is

$$H(\theta|\hat{\theta}_N) = -\sum_{i=1}^{M} p_i \sum_{j=1}^{M} q_{ij} \log q_{ij} \tag{18}$$

The empirical mutual information between θ and $\hat{\theta}_N$ is

$$I(\theta; \hat{\theta}_N) = H(\theta) - H(\theta|\hat{\theta}_N) \tag{19}$$

This empirical mutual information depends on the data length N. The approximation accuracy to the true mutual information depends on the partition and the proper choice of d_M. The maximum empirical mutual information can be achieved when the estimator always chooses the correct region among the M partitions if the true value of θ is in that region. If the performance evaluator has additional knowledge about the construction of the estimator in an analytic form, then the asymptotic performance using uniform quantization can also be evaluated [6].

For the estimation of a random process, i.e., a filtering problem, the PE can choose a sequence of unknown parameters $\theta^S = \{\theta_1, ..., \theta_S\}$ at $t_1, ..., t_S$ as the representative points (samples) of the process and evaluate the distortions at those times based on the M partitions of θ_i $(i = 1, ..., S)$. Assuming that at any time t_i the M partition of θ_i is d_M bounded, the *empirical average mutual information* is approximated by

$$I(\theta(t), \hat{\theta}_N(t)) \approx \frac{1}{S} I(\theta^S; \hat{\theta}_N^S) \tag{20}$$

In the above definition, we also assume that a data sequence of length N is generated at each time t_i $(i = 1, ..., S)$. Clearly, the empirical mutual information between θ_i and $\hat{\theta}_i$ depends on N. If $\{\theta_1, ..., \theta_S\}$ are independent, then we have

$$I(\theta(t), \hat{\theta}_N(t)) \approx \frac{1}{S} \sum_{i=1}^{S} I(\theta_i; \hat{\theta}_{i,N}) \tag{21}$$

To have a performance measure independent of N, we will focus on the asymptotic information gain as N increases. In a classification problem, the information gain from one additional observation is $I(X; \hat{X}_{N+1}) - I(X; \hat{X}_N)$. However, as N goes to infinity, the information gain can approach to zero. Denote by ΔI_N the information gain by using $N + 1$ observations instead of N observations. For an estimation problem, we have

$$\Delta I_N = I(\theta; \hat{\theta}_{N+1}) - I(\theta; \hat{\theta}_N) \tag{22}$$

For a filtering problem, we have

$$\Delta I_N = I(\theta(t); \hat{\theta}_{N+1}(t)) - I(\theta(t); \hat{\theta}_N(t)) \tag{23}$$

The mutual information can be computed using the empirical mutual information with d_M bounded partition. If there exists a value β such that

$$0 < \lim_{N \to \infty} N^{-\beta} \Delta I_N = C(\beta) < +\infty \tag{24}$$

then we define the *asymptotic information rate* as β and the gain as $C(\beta)$.

For an unknown parameter θ with Gaussian prior distribution being observed under additive white Gaussian noise, we have $\beta = 1$ and and $C(\beta) = 0.5$ [8]. For a target with white noise acceleration motion being observed by N sensors with position measurements under additive white Gaussian noises independent across sensors, in the steady state, the centralized estimator yields $\beta = 0.75$ while the distributed estimator using

track-to-track fusion without any feedback has $\beta = 0$ [5]. Clearly, a larger value of β indicates a better rate of information gain in the asymptotic regime. If two algorithms have the same rate β, the one with a larger $C(\beta)$ is expected to have a better performance for large N. To estimate the asymptotic information rate empirically, the PE needs to compute the increase of the mutual information due to an additional observation as a function of N and uses the log plot to find the slope within a certain range for large N. We will elaborate this with additional illustrative examples in the next section.

4 Illustrative Examples

Example 1 (classification) [14]: Consider a classification problem where the PE provides input $\mathbf{x}$ to the AD and the AD outputs $y(\mathbf{x})$ so that the PE can compare y with the true class value t. The PE used 100 testing samples to evaluate three classifiers A–C with the confusion matrix as below. We can see that both classifiers A and B have the same error rate of 10% and classifier C has a larger error rate of 12%. However, classifier A simply guesses that the outcome is 0 for all cases while classifier B makes no error when declaring $y = 0$ and has a 50% chance being correct when declaring $y = 1$ as opposed to the prior probability $P(t = 1) = 0.1$. Clearly, the PE knows that the mutual information from classifier B is larger than that from classifier A. Using normalized error to information ratio measure, no matter what α_N the PE chooses, classifier A will always be the worst while classifier B is better than classifier C.

Classifier A

y	0	1
t=0	90	0
t=1	10	0

Classifier B

y	0	1
t=0	80	10
t=1	0	10

Classifier C

y	0	1
t=0	78	12
t=1	0	10

Next, we assume that the classifier can output a '?' indicating that it is not sure whether the input should belong to any of the two classes. The PE wants to compare classifiers D and E with 100 testing samples and the confusion matrix is shown as below. Both classifiers D and E have 6% error rate and 11% rejection rate ('?'), however, one should not conclude that they have the same classification performance. Classifier E is just the classifier C in disguise: When C declares $y = 1$, E will toss a coin with equal chance declaring $y = 1$ and $y =$?. Classifier D is more informative than classifier B: it makes no error when declaring $y = 0$ and has a 60% chance being correct when declaring $y = 1$. Again, using normalized error to information ratio, no matter what α_N the PE chooses, classifier D is always better than classifier E. It is in line with our intuition that classifier B performs better than classifier C.

Classifier D

y	0	?	1
t=0	74	10	6
t=1	0	1	9

Classifier E

y	0	?	1
t=0	78	6	6
t=1	0	5	5

Note that the most informative classifier to the PE does not necessarily provide the best classification accuracy. However, the information theoretic measure is meaningful especially when the AD does not provide the PE the statistical model that the classification algorithm is built upon but the classifier itself.

Example 2 (hypothesis testing): Consider a binary detection problem with the observation sequence given by $z_i = \theta + w_i$, $i = 1, ..., N$. The signal θ is modeled by the following two simple hypotheses H_0: $\theta = 0$ vs. H_1: $\theta = 1$. The noise sequence is white and w_i follows the double exponential distribution with the pdf

$$p(w) = \frac{1}{2} e^{-|w|} \tag{25}$$

One detector uses the likelihood ratio test that computes the test statistic

$$T_{1N} = \sum_{i=1}^{N} u_i \tag{26}$$

where

$$u_i = \begin{cases} 1 & z_i > 1 \\ 2z_i - 1 & 0 \le z_i \le 1 \\ -1 & z_i < 0 \end{cases} \tag{27}$$

Another detector uses the sign test

$$T_{2N} = \sum_{i=1}^{N} \text{sign}(z_i - 0.5) \tag{28}$$

and compare with the threshold 0. The sign test is not optimal but it only assumes that the noise pdf is symmetrical around zero. The performance evaluator has equal prior probability on the two hypotheses and wants to compare the two detectors using the information theoretic measure. Figure 1 shows the empirical mutual information as a function of the total observations N. We can see that the likelihood ratio detector has a slightly better performance than the sign detector and the performance gap decreases as N increases. When N approaches infinity, both detectors achieve the maximum information of 1 which implies no detection error. Note that, under small N (low SNR regime), the sign detector has the mutual information close to that of the likelihood ratio detector. This is consistent with the standard analysis that the sign detector is locally optimal [9]. In this example, we can see that the sign detector does not lose performance by much compared with the optimal likelihood ratio detector.

Example 3 (parameter estimation): A coin has a probability p of coming up heads which is unknown. The performance evaluator tosses the coin N times and

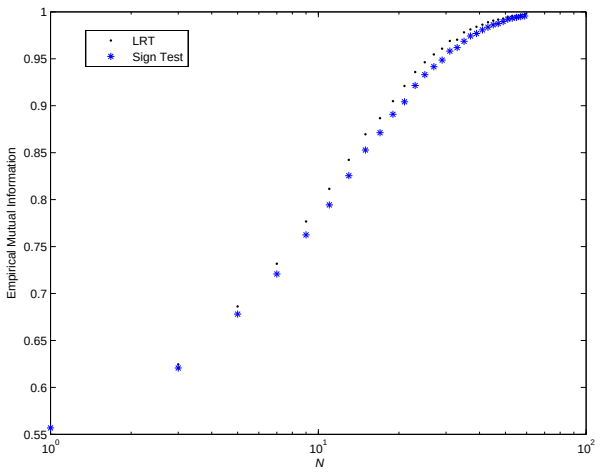

Figure 1: Comparison of the empirical mutual information between the likelihood ratio detector and the sign detector.

M heads have occurred. If the performance evaluator has a uniform prior (corresponding to a Beta distribution with parameters $B(1,1)$) on p, then the posterior of p is $B(M+1, N-M+1)$ where $B(m,n)$ denotes Beta distribution with pdf

$$f(x) = \frac{\Gamma(m+n)}{\Gamma(m)\Gamma(n)} x^{m-1}(1-x)^{n-1} \tag{29}$$

The mutual information is the difference of the differential entropy between the prior and the posterior on p. This Bayesian procedure requires that the performance evaluator has the knowledge of the likelihood function and the inference on p is summarized by the whole posterior density.

If the performance evaluator does not have the complete knowledge of the likelihood function and the algorithm developer only provides a point estimate on p, then the posterior density can not be fully specified. In this case, the empirical mutual information is helpful for performance comparison. Let us assume that one estimator gives $\hat{p}_1 = \frac{M}{N}$ and another estimator gives $\hat{p}_2 = \frac{M+1}{N+2}$. The performance evaluator desires to have $|p - \hat{p}| \leq 0.1$ as the distortion bound. Five possible values $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ of p are used to generate N tosses to evaluate the performance of the two estimators. Figure 2 shows the empirical mutual information as a function of N for the two estimators. We can see that for small N, the maximum likelihood estimator $\hat{p}_1$ (Estimator 1) has slightly larger mutual information than the other estimator $\hat{p}_2$ (Estimator 2), which is the Bayesian predictive probability that the next toss will be a head after seeing M heads in N tosses. Note that the result does not say that $\hat{p}_1$ yields better estimation accuracy than $\hat{p}_2$, however, the performance evaluator can interpret it this way: "If p can only take five possible values, I will gain more information from estimator $\hat{p}_1$ than from estimator $\hat{p}_2$." If the performance evaluator chooses α_N=1, then the normalized error to information ratio also reveals that $\hat{p}_1$ performs better than $\hat{p}_2$. In fact, the assumptions made in $\hat{p}_1$ seems closer to the performance evaluation procedure and the results are as expected.

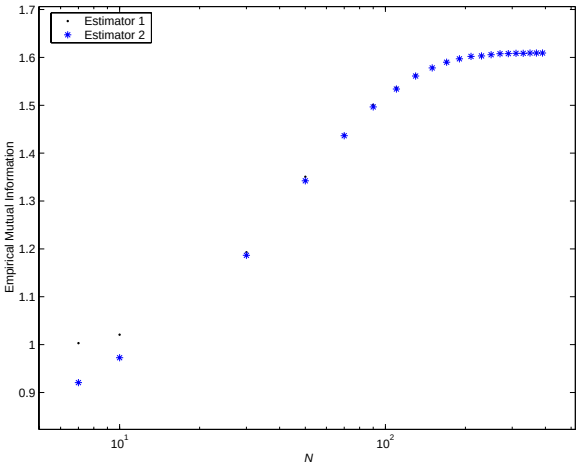

Figure 2: Comparison of the empirical mutual information between the two estimators.

Example 4 (joint classification and estimation): Consider a communication problem with the following observation equation

$$z_i = \eta x_i + n_i, i = 1, ..., N \tag{30}$$

where $x_i \in \{-1, 1\}$ is the message to be transmitted; $\eta \in (0, +\infty)$ is an unknown fading coefficient; and $n_i \sim \mathcal{N}(0, \sigma^2)$ is the additive white Gaussian noise. Given the observation sequence, one has to decode the message $\{x_i\}$ and estimate the fading parameter η jointly. In communication, one often cares only the decoding performance, however, some joint classification and estimation problems with similar setup may require to evaluate the decision and estimation errors simultaneously. Assume that the message sequence is a Markov chain with known transition probability, then the maximum likelihood estimate of both $\{x_i\}$ and η is

$$\{\hat{x}_i\}, \hat{\eta} = \arg \min_{\{x_i\}, \eta} \sum_{i=1}^{N} (z_i - \eta x_i)^2 - \log P(x_i | x_{i-1}) \tag{31}$$

which is computationally prohibitive if one has to examine all possible message sequences. An efficient method, denoted by Algorithm A, makes the decision first using

$$\hat{x_i} = \text{sign}(z_i) \tag{32}$$

and then estimate η using least squares assuming that each decoded message is correct. Alternatively, Algorithm B performs the estimation directly with

$$\hat{\eta} = \sqrt{\sum_{i=1}^{N} z_i^2 / N - \sigma^2} \tag{33}$$

and the classification of the message is made by

$$\{\hat{x}_i\} = \arg\min_{\{x_i\}} \sum_{i=1}^{N} (z_i - \hat{\eta} x_i)^2 - \log P(x_i|x_{i-1}) \quad (34)$$

using Viterbi algorithm. By applying empirical mutual information measure, we found that Algorithm B is better than A when $P(x_i|x_{i-1}) = \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}$, N and σ^2 are large enough. This confirms with our intuition that incorrect decision using the sign detector also leads to poor estimation accuracy of the fading parameter. Thus the proposed measure can also be meaningful to evaluate algorithms developed for other joint classification and estimation problems with more complex structure where classification or estimation performance alone may not be the only focus by the PE.

Example 5 (joint classification and filtering): Consider a dynamic system with state equation

$$x_k = F_j x_{k-1} + v_{k-1} \quad (35)$$

where $F_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $F_2 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$. The noise $v_k \sim \mathcal{N}(0, Q)$ is white Gaussian sequence with $Q = \begin{bmatrix} \frac{1}{3} & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix} \sigma_v^2$. The observation model is

$$z_k = H x_k + w_k \quad (36)$$

where $H = \begin{bmatrix} 1 & 0 \end{bmatrix}$ and $w_k \sim \mathcal{N}(0, \sigma_w^2)$ is white Gaussian sequence independent of v_k. We are interested in sequentially estimate the state x_k and classify the dynamic model F_j. Denote by M_k the model at time k and assume that the model sequence M_k is a Markov chain with transition probability matrix $P(M_k|M_{k-1}) = \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}$. The system is of linear Markov jump type which can be extended to handle maneuvering target tracking [2] and joint target classification and tracking [4] problems.

Clearly, the conditional density of the state x_k is a Gaussian mixture

$$p(x_k|Z^k) = \sum_{l=1}^{2^k} p(x_k|M^{k,l}, Z^k) P(M^{k,l}|Z^k) \quad (37)$$

with exponentially increasing components. The probability $P(M^{k,l}|Z^k)$ of a model history can be obtained recursively using Bayes' formula [1]. Consider Algorithm A that always keeps 16 model history sequences with the largest probabilities, discards the rest of the sequences, and renormalizes the probabilities. It can be interpreted as multiple hypothesis tracking (MHT) [2] with hypothesis pruning. Another Algorithm B combines the histories of models and keeps only the possible models in the last two sampling periods. Algorithm B requires 4 filters to operate in parallel and is called the generalized pseudo-Bayesian of order 2 (GPB2) [1]. Algorithm C uses interacting multiple model (IMM) [1] and makes decision sequentially according to the approximate model probability $P(M_k|Z^k)$. We assume that $\sigma_w^2 = \sigma_v^2 = 1$ and compute the asymptotic information rate for Algorithms A–C. Based on the approximate slope from a state and model sequence of length 200, we obtained $\beta_A = 0.112$, $\beta_B = 0.088$ and $\beta_C = 0.084$. This is in line with the existing results reported on the state estimation accuracy [1]: IMM has comparable tracking error to GPB1 (but with less computational requirement) and MHT with hypotheses of longer time history can further improve the tracking accuracy. Interestingly, Algorithm B has slightly smaller average classification error on the choice of the model at each time than Algorithm C. This is also reflected from a small difference in their asymptotic information rates.

5 Relation to Other Performance Measures

Mutual information was originally proposed to characterize the capacity of a communication channel [6]. Its extension to evaluate an estimator is usually based on the Fisher information, which is related to the Cramer-Rao lower bound of the mean square estimation error [1]. Another connection between the mutual information and mean square estimation error in Gaussian channel was discussed in [8]. However, these information theoretic measures are *algorithm independent*. For evaluating the quality of a classifier, a complete error rate vs. rejection rate curve is usually generated for each classifier. Some people use the area under this curve to rank different classifiers [14]. An alternative metric to classifier ranking in the Neyaman-Pearson paradigm was proposed in [16]. There exists abundant performance measures for the evaluation of estimation algorithms [17, 3, 13]. However, practitioners often use the mean square estimation error to rank estimation and filtering algorithms. In addition, if an algorithm also provides its self assessment on the mean square estimation error, the evaluation of this additional information often requires the credibility test [11, 12]. There is no comprehensive measure of the credibility of an estimator except the noncredibility index (NCI) proposed in [10]. However, NCI can not be easily extended to evaluate the credibility of a classifier even when the classifier can provide its self-assessment of the classification accuracy in terms of the confusion matrix. There are inherent difficulties in evaluating the performance of joint classification and estimation algorithms. Estimation error and classification error are often incompatible and the performance evaluator may not have access to the precise description of the algorithm but its behavior via testing examples. Thus empirical mutual information and asymptotic information rate are important

indicators for the performance evaluator to meaningfully judge an algorithm's quality or compare the performance among different algorithms through carefully controlled scenarios. There is no need to treat the decision and estimation problems separately. On the other hand, we do not intend to replace the existing performance measures or metrics with information theoretic measures, but to compliment them in the algorithm evaluation and comparison problems of practical interest.

6 Discussions and Conclusions

In this paper, we studied the performance evaluation using several newly derived information theoretic measures, namely, the empirical mutual information, normalized error to information ratio, and the asymptotic information rate for classification, estimation and filtering problems. They serve as a guideline for designing a practical procedure to measure the performance of different algorithms with limited knowledge of the parametric model that an algorithm developer is based upon. Several practical examples including joint decision and estimation are used for algorithm comparison and for gaining the insight on some inherent difficulties of algorithm ranking. In most cases, information theoretic measures do rank the performance of the candidate algorithms properly even in the joint classification and estimation problem where classification or estimation accuracy alone does not provide the complete picture of algorithm performance.

Acknowledgement

This work was supported in part by the US Air Force under contracts FA9453-09-C-0175 and FA8650-09-M-1552, Army Research Office under contract W911NF-08-1-0409, and Louisiana Board of Regents under contract NSF(2009)-PFUND-162. The authors are grateful to the anonymous reviewers for helpful comments and to Dr. Xiaokun Li at DCM Research Resources, LLC, and Dr. V. P. Jilkov at University of New Orleans for stimulating discussions.

References

[1] Y. Bar-Shalom, X. Rong Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation: Theory, Algorithms and Software*, Wiley, 2001.

[2] Y. Bar-Shalom, and X. Rong Li, *Multisensor, Multitarget Tracking: Principles and Techniques*, YBS Publishing, 1995.

[3] S. Blackman, and R. Popoli, *Design and Analysis of Modern Tracking Systems*, Artech House, 1999.

[4] S. Challa, and G. W. Pulford, "Joint Target Tracking and Classification Using Radar and ESM Sensors", *IEEE Trans. Aerospace and Electronic Systems*, 37(3), pp. 1039–1055, Jul. 2001.

[5] H. Chen, and X. R. Li, "On Track Fusion with Communication Constraints", Proc. of *10th International Conference on Information Fusion*, Québec, Canada, July 2007.

[6] T. M. Cover, and J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, 1991.

[7] R. Gupta, and A. O. Hero, "High-Rate Vector Quantization for Detection", *IEEE Trans. on Information Theory*, 49(8), pp. 1951–1969, Aug. 2003.

[8] D. Guo, S. Shamai and S. Verdu, "Mutual Information and Minimum Mean-Square Error in Gaussian Channels", *IEEE Trans. Information Theory*, vol. 51, pp. 1261–1282, April 2005.

[9] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory*, Prentice Hall, 1998.

[10] X. R. Li and Z.-L. Zhao, "Measuring Estimator's Credibility: Noncredibility Index", *Proc. 2006 International Conf. on Information Fusion*, Florence, Italy, July 2006.

[11] X. R. Li and Z.-L. Zhao, "Testing Estimator's Credibility – Part I: Tests for MSE", *Proc. 2006 International Conf. on Information Fusion*, Florence, Italy, July 2006.

[12] X. R. Li and Z.-L. Zhao. "Testing Estimator's Credibility – Part II: Other Tests" *Proc. 2006 International Conf. on Information Fusion*, Florence, Italy, July 2006.

[13] X. R. Li, and Z.-L. Zhao, "Evaluation of Estimation Algorithms – Part I: Incomprehensive Performance Measures", *IEEE Trans. Aerospace and Electronic Systems*, 42(4), Oct. 2006.

[14] D. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge Press, 2003.

[15] A. Papoulis, and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*, McGraw Hill, 2002.

[16] C. Scott, "Performance Measures for Neyman-Pearson Classification", *IEEE Trans. Information Theory*, 53(8), pp. 2852–2863, Aug. 2007.

[17] E. Waltz, and J. Llinas, *Multisensor Data Fusion*, Artech House, 1990.